THE STUDY OF SPEECH/PAUSE DETECTORS FOR SPEECH ENHANCEMENT METHODS

Pavel SOVKA & Petr POLLÁK
E-mail: sovka@feld.cvut.cz, pollak@feld.cvut.cz
Czech Technical University, Faculty of Electrical Engineering, K331
Technická 2, 166 27 Prague 6, CZECH REPUBLIC
Fax: (+42 2) 2431 0784

ABSTRACT

Basic principles of various adaptive algorithms for speech detection in a noise and their behaviour under real car noise conditions are described. Energy, spectral, cepstral, and coherence detectors are compared. All these algorithms are suitable for real time implementation with one or two microphones. High probability of correct speech/pause detection can be obtained even if signal to noise ratio is low and noises are highly nonstationary.

1. INTRODUCTION

Speech/pause detectors are the limiting parts of systems for the suppression of additive noises in speech, because the quality of the detector determines the performance of the whole noise suppression system. If the speech/pause decision is not correct then speech echoes and residual noises are present in enhanced speech. Information about speech activity is need not only for an estimation of background noise characteristics but also for time delay compensation of signals picked up by microphone array.

2. SPEECH/PAUSE DETECTION

Basic steps of speech detection can be formulated as follows:

- signal is divided into overlap segments
- a vector of chosen parameters is computed for signal segments
- a proper distance measure is used for evaluation of a difference between two segments
  - if the distance between given segment and a preset threshold is evaluated then we obtain an *integral algorithm* which is able to detect the whole interval of speech activity
  - if the distance between two neighbouring segments is evaluated we obtain a *differential algorithm* which can detect borders of speech activity only

An integral algorithms require to preset a threshold, but this setting suffers from the subjectivity which is not the case of a differential algorithms. On the other hand the output from differential algorithms have to be postprocessed (e.g. integrated) to give the whole intervals of speech activity. This postprocessing introduces some errors and degrades the performance of differential algorithms.

2.1. Chosen parameters

The choice of vector parameters and the distance measure determine the type of detector. Energy, spectral density, cepstrum and coherence function were used. We studied following parameters:

Energy - computed either in the time or frequency domain, i.e.

$$E = \sum_n x^2[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} S(e^{j\Theta}) d\Theta, \qquad (1)$$

were $S(e^{j\Theta})$ is energy spectral density (ESD) of a signal $x[n]$ and Θ is normalized frequency in radians.

Energy spectral density - $S(e^{j\Theta})$ - given using the FFT or AR analysis.

Cepstral coefficients - $c[n]$ - determined by Laurent expansion of ESD

$$\ln S(e^{j\Theta}) = \sum_{k=-\infty}^{\infty} c_k e^{-jk\Theta}. \qquad (2)$$

Energy detector can be thought as the special case of cepstral detector because c_0 contains information about signal energy.

Coherence function

$$\gamma^2(e^{j\Theta}) = \frac{|S_{xy}(e^{j\Theta})|^2}{S_x(e^{j\Theta}) S_y(e^{j\Theta})}, \qquad (3)$$

where $|S_{xy}(e^{j\Theta})|^2$ is cross spectrum of two input signals $x[n]$ and $y[n]$, $S_x(e^{j\Theta})$ and $S_y(e^{j\Theta})$ are ESD of $x[n]$ and $y[n]$ respectively. Coherence function approaches to 1 when $x[n]$ and $y[n]$ are clean real speech signals. On the other hand coherence function falls to 0 when $x[n]$ and $y[n]$ are uncorrelated noises. Because most of car noises seems to be uncorrelated coherence function contains some information about speech activity. Disadvantage is that two microphones are required.

2.2. Distance measures

We used two types of distance measures:

$$d_S = \frac{1}{2\pi}\int_{\Theta_{min}}^{\Theta_{max}} |\ln S(e^{j\Theta}) - \ln S'(e^{j\Theta})| d\Theta \qquad (4)$$

where Θ_{min} and Θ_{max} were set experimentally. This distance is L1 norm of spectral distance between two signal segments. If $S'(e^{j\Theta})$ is ESD of a background noise then eq.(4) describes integral spectral detector, if $S'(e^{j\Theta})$ is ESD of a previous segment then eq.(4) leads to the differential spectral detector.

$$\begin{aligned} d_{cep}^2 &= \frac{1}{2\pi}\int_{-\pi}^{\pi} |\ln S(e^{j\Theta}) - \ln S'(e^{j\Theta})|^2 d\Theta \\ &= \sum_{k=-\infty}^{\infty} (c_k - c'_k)^2 \qquad (5) \end{aligned}$$

This is L2 norm of the difference between two cepstral vectors. If c'_k are cepstral coefficients of a background noise then eq.(5) leads to an integral cepstral detector. If $c_k = c_k[n]$ and $c'_k = c_k[n-1]$ are cepstral coefficients of two adjacent signal segments at the time n and $n-1$ respectively, then the eq.(5) can be used for the construction of the differential cepstral detector.

2.3. Thresholds updating

Threshold updating can be described as follows:

- Let $\mathbf{V}[n]$ is the vector of length M of specified parameter, i.e. $M = 1$ for energy, M is number of cepstral coefficients, or M is segment length for the case of $S(e^{j\Theta})$ or $\gamma(e^{j\Theta})$.

- Then smoothed version $\mathbf{V_S}[l]$ of $\mathbf{V}[l]$ is computed in nonspeech activity

  $$\mathbf{V_S}[l+1] = p\mathbf{V_S}[l] + (1-p)\mathbf{V}[l] \qquad (6)$$

  Parameter p corresponds the equivalent window length of $1/p$ - it means $1/p$ segments because $\mathbf{V}[l]$ is computed for one signal segment at time l. This smoothed process is expected to be short-time.

- chosen distance $dist[n]$ between $\mathbf{V}[l]$ and $\mathbf{V_S}[l]$ is evaluated using equations (4), (5), or (8).

- mean value and standard deviation of computed distance $dist$ is estimated using exponential averaging with the equivalent window length of $1/q$. This averaging is expected to be long-time.

- threshold Thr is then established as

  $$Thr = \text{mean}\,(dist) + \alpha \cdot \text{std}\,(dist) \qquad (7)$$

  where α was found to be in the range from 1 to 2 according the signal to noise ratio (SNR) and type of detector.

2.4. Detector specifications

2.4.1. Energy detector

The value E from eq.(1) is compared with the threshold E_t defined as $E_t = 1.5\,E_b$ where E_b is the background noise energy updated according eq.(6) in nonspeech activity, i.e. while $E < E_t$. The threshold could be defined according eq.(7) too.

2.4.2. Spectral integral detectors

This detector using either measure d_S (4) with adaptive threshold according eq.(7) or the statistics [1]

$$D = [N_\Theta(\frac{2}{n_1} + \frac{2}{n_2})]^{-\frac{1}{2}} \sum_{i=1}^{N_\Theta} \log\frac{S(e^{j\Theta})}{S'(e^{j\Theta})}, \qquad (8)$$

which has a standardized normal distribution, n_1 and n_2 stand for degrees of freedom of $S(e^{j\Theta})$ and $S'(e^{j\Theta})$ respectively, and N_Θ is the number of frequency bands. Equation (8) provides a basic to test the hypothesis that $S(e^{j\Theta}) = S'(e^{j\Theta})$. The region of acceptance for the hypothesis test is

$$[-z_{\alpha/2} \le D \le z_{\alpha/2}], \qquad (9)$$

where α is the level of significance for the test and $z_{\alpha/2}$ is $100\frac{\alpha}{2}$ percentage point. Our goal was to find an optimal values n_1, n_2, and N_Θ. If eq.(8) is realized by the FFT of size N then maximum value of $N_\Theta = N/2$.

2.4.3. Cepstral integral detectors

The suggestions of these algorithms were motivated by [2], [3], [4], [5], [7], [10], [11]. Eq.(5) is approximated in this case by

$$\Delta c = 4.3429\sqrt{(c_0 - c'_0)^2 + 2\sum_{k=1}^{p}(c_k - c'_k)^2}\,. \qquad (10)$$

This algorithm were used in one step or two step version. One step algorithm computes with Δc only while two step one works with smoothed distance Δc_m.

2.4.4. Cepstral differential detectors

This detector computes a first order differential log spectrum

$$\frac{\partial \ln S(e^{j\Theta})}{\partial t} = \sum_{k=-\infty}^{\infty} \frac{\partial c_k(t)}{\partial t} e^{-jk\Theta}, \qquad (11)$$

where $c_k(t)$ stands for $k-$th cepstral coefficients at time t. This equation can not be used for the time-sampled cepstral sequence. That is why we studied various approximations of the time derivation. One of these possibilities is *backward difference*, i.e.

$$\delta_k^{(1)}(\tau) = \frac{\partial c_k(t+\tau)}{\partial t}\Big|_{t=0} \approx c_k[n] - c_k[n-1], \qquad (12)$$

which is inherently noisy. It is possible to use *symmetric difference* which represents reasonable choice between complexity and performance or *polynomial approximation* which gives the best detection performance for lower orders. The speech detection was performed using a modified differential spectral distance. Following eq.(13) is the special case of eq.(5) for $\delta_k^{(1)} = c_k[n] - c_k[n-1] = c_k - c'_k$.

$$d_d^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |\frac{\partial \ln S(e^{j\Theta},t)}{\partial t}|^2 d\Theta. \approx \sum_{k=-\infty}^{\infty} (\delta_k^{(1)})^2 \qquad (13)$$

2.4.5. Coherence integral detector

This part of study was motivated by paper of [6]. This detector uses two input signals $x[n]$ and $y[n]$ to estimate coherence function in eq.(3). Then this function is integrated

$$M_\gamma = \int_{\Theta_{min}}^{\Theta_{max}} \gamma^2(e^{j\Theta})d\Theta. \qquad (14)$$

Parameter M_γ is compared to:

- fixed threshold M_t which value was varied in interval $(0.2, 0.4)$
- threshold M_{ta} updated in nonspeech activity established according eq.(7)

Bounds Θ_{min} and Θ_{max} together with segments length determines the detector behaviour.

3. REALIZATION

All algorithms were developed under constrained that the length of segments has to be 256 samples with 50% overlapping and sampling frequency $11025Hz$. Energy and cepstral algorithms fully satisfy these requirements.

Smoothing process of spectral and coherence detector requiring high number of degrees of freedom was based on Welch method. The FFT size was chosen 32 for spectral detectors. Other parameters of statistical spectral detector were set $n_1 \approx 70$, $n_2 \approx 90$, and $\Theta \approx 12$ with corresponding unnormalized frequency band $(350, 3800)\,Hz$[1].

Experiments confirmed that the performance of coherence detector is poor for segment length 256. That's why segments of 1024 or 2048 samples were used. In this case the FFT size is 128. The unnormalized frequency band $(400, 4000)$ Hz was used for presented coherence algorithms[2].

Detector	$P(A/s)$	$P(A/N)$	$P(A)$	$P(B)$
1.HDET	0.8540	0.5120	0.6490	0.4310
	0.0070	0.0170	0.0040	0.0080
2.CEP1	0.9700	0.6650	0.8010	0.6420
	0.0020	0.0200	0.0090	0.0160
3.CEP2	0.9620	0.7670	0.8630	0.7340
	0.0020	0.0210	0.0070	0.0160
4.DF1	0.9660	0.6970	0.8280	0.6710
	0.0020	0.0130	0.0060	0.0110
5.PSD	0.8420	0.6640	0.7710	0.5600
	0.0140	0.0150	0.0090	0.0190
6.PSDA	0.9160	0.5540	0.6670	0.5060
	0.0180	0.0270	0.0270	0.0280
7.COHA	0.8520	0.6310	0.7470	0.5370
	0.0090	0.0180	0.0090	0.0170
8.COHF	0.9810	0.4760	0.5870	0.4670
	0.0010	0.0180	0.0160	0.0170

Tab. 1: Average values and standard deviations of classification parameters for experiments with selected detectors and with SNR $= 0dB$.
1.HDET - Harrison energy detector,
2.CEP1 - One step integral cepstral detector,
3.CEP2 - Two step integral cepstral detector,
4.DF1 - Differencial cepstral detector (backward dif.),
5.PSD - Statistical spectral detector,
6.PSDA - Spectral detector with adaptive threshold,
7.COHA - Coherence detector, adaptive threshold,
8.COHF - Coherence detector, fixed threshold.

4. CLASSIFICATION

We used following objective criteria (based on the computation of correct detection rates) for the comparison of different algorithms:
- *correct speech detection rate* - $P(A/S)$
- *correct nonspeech detection rate* - $P(A/N)$.

Another possibility is the using of global criteria:
- *correct detection rate* - $P(A)$
- *speech/nonspeech resolution factor* - $P(B)$

defined as

$$P(A) = P(A/S)P(S) + P(A/N)P(N),$$
$$P(B) = P(A/S)P(A/N)$$

where $P(S)$ and $P(N)$ are rates of speech and pauses in the processed signal. While first and second criteria show how a detector determines speech or noise activity only, third and fourth criteria show the global detector performance.

5. EXPERIMENTS

Detectors were tested under different noisy conditions - SNR $= 0 \div 20\,dB$. Speech signals were manually labeled and mixed with real car noise. A rough information about detectors behaviour is given by mean values and standard deviations of all realized experiments for each criterion described in preceeding text. Tab.1 and fig.1 show some of these results. More detailes of achieved results will be presented.

6. CONCLUSIONS

All described detectors are suitable for real time implementation. We focused to AR-based cepstral de-

[1] Non-linear frequency warping was used for spectral and coherence detectors too but the results were similar to linear frequency scale [8].

[2] AR based coherence detector was developed but results in the speech/pause resolution were worse than results of FFT coherence detector. Moreover this type of detector require long signal segments and very high order of AR model and it is more sensitive to noise than FFT based one.

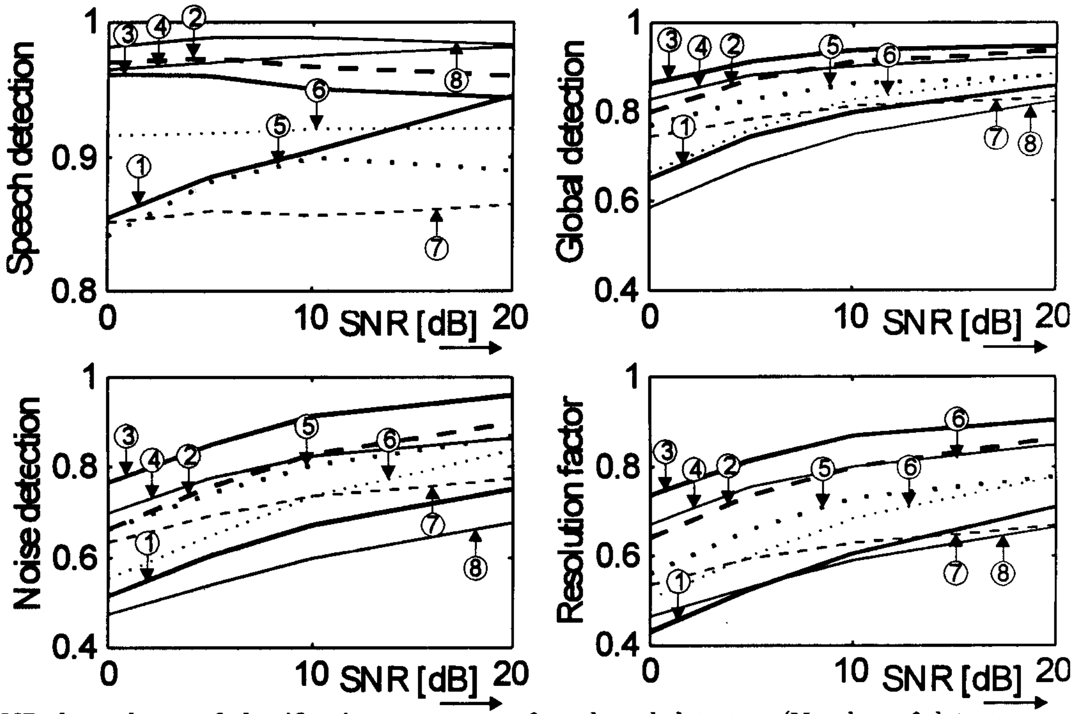

Fig. 1: SNR dependance of classification parameters for selected detectors.(Number of detectors corespond to the marks in the presented table.)

tectors only because the results are better than for FFT-based ones because of following reasons:

- lower noise fluctuation in AR-based cepstra,
- greater sensitivity of AR-based cepstra to speech activity,
- threshold can be set to lower value because more smoothed estimations of background noise is obtained.

Because the segment length of coherence detector is greater than segment length of other types of detectors, coherence detector gives global long-term information about speech activity while cepstral detectors give detail description of speech activity. Presented coherence detectors are the simplest versions of this type of algorithms and other modifications are under development.

The best results gives two steps integral cepstral detector with nonlinear smoothing described in [9].

REFERENCES

[1] J. S. Bendat and A. G. Piersol. *Random Data Analysis And Measurment Procedure.* Willey-Interscience, New York, 1971.

[2] A. H. Gray, Jr. and J. D. Markel. "Distance measures for speech processing". *IEEE Trans. on Acoustics, Speech, and Signal Processing,* ASSP-24(5), pp.380-391, October 1976.

[3] J. A. Haigh and J. S. Mason. *A voice activity detector based on cepstral analysis.* In EUROSPEECH'93 - Proceedings of the 3rd European Conference on Speech, Communication, and Technology, pp.1103-1106, Berlin, September 1993.

[4] W. A. Harrison, J. S. Lim, and E. Singer. "A new application of adaptive noise canceller." *IEEE Trans. on Acoustics, Speech, and Signal Processing,* ASSP-34(1), pp.21-27, February 1986.

[5] J.-C. Junqua, H. Wakita, and H. Hermansky. "Evaluation and optimization of perceptually-based ASR front-end." *IEEE Trans. on Audio and Speech Processing,* 1(1), pp.39-47, Jan 1993.

[6] R. Le Bouquin and G. Faucon. "Using the coherence function for noise reduction." *IEE Proceedings - I,* 139, pp.276-280, June 1992.

[7] A. Le Floc'h, R. Salami, B. Mouy, and J.-P. Adoul. *Evaluation of linear and nonlinear subtraction metods for enhancing noisy speech.* In Speech Processing in Adverse Conditions, pp.131-134, Cannes-Mandelieu (France), November 1992.

[8] P. Pollák and P. Sovka. Spectral subtraction and speech/pause detection. Research report R1-95, CTU - Faculty of Electrical Engineering, Prague, 1995.

[9] P. Pollák, P. Sovka, and J. Uhlíř. *Cepstral speech/pause detectors.* In Proceedings of IEEE Workshop on Nonlinear Signal and Image Processing, Neos Marmaras, Greece, June 1995.

[10] L. R. Rabiner and B-H Juang. *Fundamentals of Speech Recognition.* Murray Hill, New Yersey, USA, 1993.

[11] L. R. Rabiner and M. R. Sambur. "Application of an LPC distance measure to the voiced-unvoiced-silence detection problem." *IEEE Trans. on Acoustics, Speech, and Signal Processing,* ASSP-25(4),pp.338-343, August 1977.